PAPER • **OPEN ACCESS**

Salt Commodity Data Clustering Using Fuzzy C-Means

To cite this article: M. Fuad *et al* 2022 *J. Phys.: Conf. Ser.* **2406** 012025

View the article online for updates and enhancements.

You may also like

- Analysis on the Country Differences of CSR of Multinational Corporations Based on Fuzzy C-Means Clustering
GuoHang Cui
- Using fuzzy C-means Clustering to Identify Heavy Metal Polluted Soil in a Certain Area of Shanghai
Yuhan Xu, Luolei Zhang, Peng Yu et al.
- Determination System Of Food Vouchers For the Poor Based On Fuzzy C-Means Method
D R Anamisa, M Yusuf and M A Syakur

This content was downloaded from IP address 140.0.123.248 on 23/12/2022 at 10:43

Salt Commodity Data Clustering Using Fuzzy C-Means

M. Fuad[1,*], E. M. S. Rochman[1], A. Rachmad[1]

[1]Faculty of Engineering, University of Trunojoyo, Madura, Bangkalan, Indonesia

Corresponding author's e-mail: *fuad@trunojoyo.ac.id

Abstract. Indonesia has a sea area of about 96,079.15 km^2 rich in natural resources. The advantage is in the form of abundant natural resources in the sea including fish and salt. Several regions such as Cirebon, Indramayu, Rembang, and Madura contribute to the salt pond commodity which is extremely valuable for Indonesia. These resources need to be monitored and inventoried appropriately. Local people are difficult and increasingly cornered to compete in the salt trade. For the strategic commodity group, by maintaining the stability of the salt commodity in the community with its irreplaceable function, the salt trade system is regulated. The purpose of this research is to use the Fuzzy C-Means Clustering method for grouping national salt commodities. This study used a test with 10 clusters for the distribution of training and testing data purposes. The research trials were carried out using the mean imputation method and grouping using the Fuzzy C-Means Clustering method. The test results of the Fuzzy C-Means Clustering method using the Silhouette Coefficient method show that the Fuzzy C-Means Clustering method with the calculation of the closest distance using Manhattan Distance is the best choice in research with the results obtained in the Silhouette Coefficient assessment is at a value of 0.274880 with a value of k = 2.

1. Introduction

Among the many commodities in Indonesia, salt has a strategic value. The supply and demand of salt commodity tend to be equal and the value has been increasing since 2015. This large supply of salt from Indonesia is supported by a fairly wide coastline. On the other hand, the demand for salt in large quantities is triggered not only from daily household consumption needs but also for industrial purposes. With a coastline of more than 108,000 km, Indonesia has great potential as the largest salt producer in the world. However, it is unfortunate that not every area of coastline can be utilized for salt production to support the supply of the salt industry needs [1]. In 1995 and 1999, the national household coverage in Indonesia with iodized salt (5 ppm) was informed to be 78.2% and 81.5%, respectively [2]. Among Indonesian schoolchildren between 1982 and 2003, there has been a decline in the prevalence of goiter from 29% to 11% [3]. The average urine iodine concentration is a readable 229 g/L according to the results of a national survey in 2003 [4]. This value is well above the minimum standard for a median iodine content of 100 g/L in urine as advised by the International Council for the Control of Iodine Deficiency Disorders [5].

The Indonesian salt industry requires the empowerment of salt production. This condition can be recognized based on the ratio of salt demand to production which is not balanced nationally [6] – [9]. From 2015 to 2019, salt is produced at a low national average value. This production value tends to fluctuate. The scarcity of salt commodities in Indonesia occurred in 2016 due to national harvest failure. However, the low national salt production does not affect the national salt demand. Instead, it continues to increase every year. The government decided to import salt to feed the national demand for salt. The process of the salt production system with graded crystallization can be seen in Figure 1.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

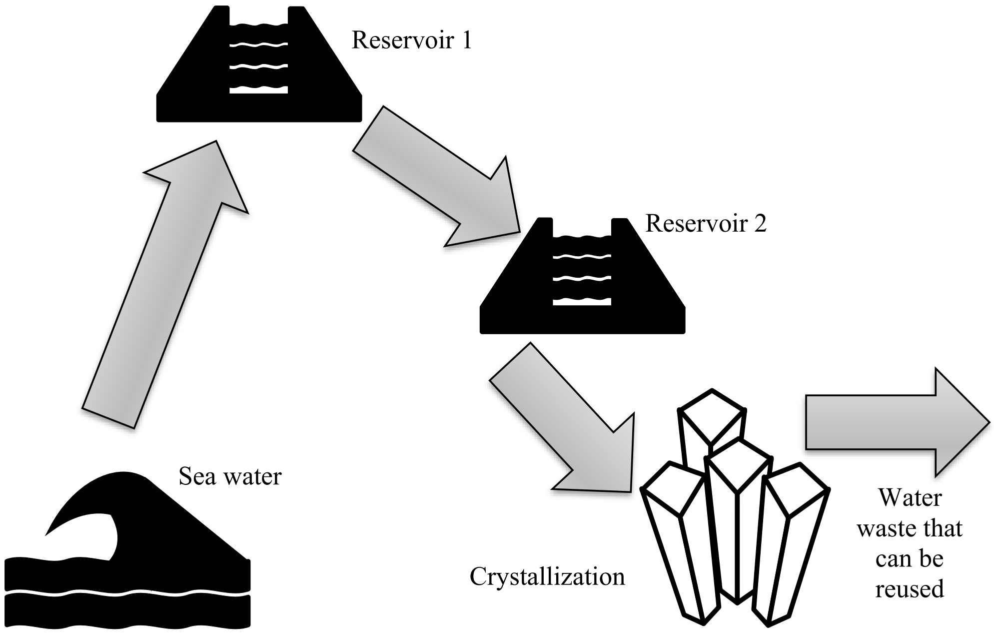

Figure 1. Salt production process.

Iodine deficiency can cause disorders including goiter, cretinism, hypothyroidism, decreased intellectual capacity, impaired physical development, increased perinatal and infant mortality and mental retardation. Iodine deficiency is estimated to affect about 36.5% of children worldwide. Iodine status can be improved by implementing universal iodized salt as a major public health strategy. The percentage of access to iodized salt by households from 5-10% increased to 68% from 1990 to 1999. This research has an urgency to develop data clustering methods to support salt trade governance. For this reason, this study proposes a novelty in the form of using the Fuzzy C-Means method in classifying salt commodities based on the Silhouette Coefficient.

2. Methodology

Data Mining is one of the important steps of the "Knowledge Discovery from Database (KDD)" process. This fairly young and interdisciplinary field of computer science is a process that tries to find interesting but hidden patterns in large data sets. According to Deepak Sinwar and Rahul Kaushik in [10], the process flow in data mining is shown in Figure 2.

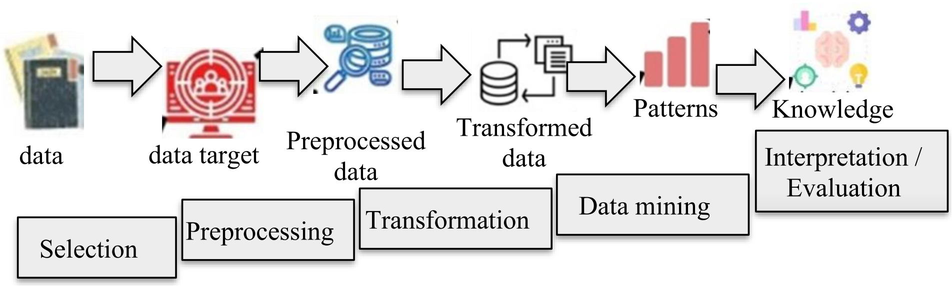

Figure 2. Data mining flowchart.

Cluster analysis is one of the most important research directions in the field of data mining. In comparison with other mining, previously the grouping of data without knowledge can complete the classification. Some parts of the clustering algorithm have types based on the model partition. The process of generating a collection of similar objects by separating different objects and collecting the same objects is known as a clustering algorithm. Data objects with the same characteristics are gathered in one cluster. Objects in one cluster have different characteristics when compared to objects from other clusters. Objects are expected to be as close together as possible within the cluster in this clustering task. The first cluster tends to be a sample or data point. Cluster aggregation that is not convergent can be caused by a random selection of sample center points. In unsupervised learning, cluster analysis is based on grouping data that have similarities [11].

The core goal of this mining process is to transform and extract information into understandable structures from data sets for further use [10].

- Classification– is the task of generalizing a well-known structure to be implemented to new data for which there is no classification. For example, classifying records based on the attribute 'class'. Prediction and Regression are also considered as part of the classification method.
- Clustering– is the task of finding groups based on the similarity of data items within the cluster and differences outside the cluster on the other side of the data set. Anomaly detection (Outlier/change/deviation detection) is also considered part of the clustering technique. This step is generally used to identify unusual/abnormal records or data errors, which can sometimes be of interest. In either case, outliers may require further investigation and processing.
- Association rule mining (Dependency modeling) – This is the task of finding interesting associations between different attributes of a dataset. Associations are generally based on new interesting but hidden patterns.

2.1. Preprocessing

To support the mining process, a preprocessing stage is needed to compile the data so that it can be used. Missing value data and data transformation are handled at this stage.

2.1.1. Cleaning and Imputation

The condition where there are void or partial values in the data is known as missing values [12]. Missing values can be handled by applying a simple statistical method known as the mean value imputation technique. In this method, any missing value is substituted with a reasonable estimate before being included in the total existing value [13].

2.1.2. Transformation Data

Data mining can be executed after the data is changed to suit the needs through the data transformation process. Discretization and normalization are several techniques for data transformation. Accurate and easier results can be obtained through discretization [14]. Continuous attribute values are converted to a finite number of intervals in Discretization. This process is continued by changing and associating each interval with a discrete numeric value. Before starting the grouping, the Min-Max data normalization was performed. This process performs a linear transformation on the raw data [15]. The normalization process describes the value of each variable to the same range, namely [0,1]. Min-Max normalization is expressed as follows:

$$X_n = \frac{X_0 - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where:
X_n = normalized data
X_0 = original data to be normalized
X_{min} = minimum value
X_{max} = maximum value

2.2. Mining Process

2.2.1. Fuzzy C-Means Clustering

Data clustering permits objects with similar characteristics to be in the same group. This step is needed to facilitate further processing. Identification of part families for mobile fabrication is one of many engineering applications of data clustering. Among the data grouping algorithm, Fuzzy C-means is quite popular. It takes the number of clusters in the data to be determined to use this method. Generally, a trial-and-error process is used to find the suitable number of clusters for a specified data set. Furthermore, this process is made more difficult due to the subjective nature of deciding what constitutes a correct grouping [11].

1. Euclidean Distance
   Euclidean Distance is usually used to calculate distances in an N-dimensional vector space. It is defined as [16]. The Euclidean distance can also be used to measure the tightness or overall spread of the assemblage distribution, which can then be compared between groups of sets such as described as follows:
   $$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$
   where:
   x_i = sample data
   y_i = centroid of the-*k* cluster
   i = attribute dimensions or each data
   n = amount of data
2. Manhattan Distance
   Manhattan Distance is a way to compute the sum of the absolute difference between the coordinates of a pair of objects. The formula used is as follows [16]:
   $$d(x, y) = \sum_{i=1}^{n}|x_i - y_i| \quad (3)$$
   where:
   x_i = sample data
   y_i = centroid of the-*k* cluster
3. Minkowski Distance
   The sum of the absolute differences in the respective coordinates is used to determine the Minkowski Distance between two points in *n*-dimensional real vector space with a fixed Cartesian coordinate system. The Minkowski distance between vectors $\vec{a}$ and $\vec{b}$ is the sum of the projected lengths of the line segments between the points to the coordinate axes [17]:
   $$d(x, y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \quad (4)$$
   where:
   x_i = sample data
   y_i = centroid of the-*k* cluster
   p = distance finder

2.2.2. Silhouette Coefficient

Silhouette Coefficient (SC) combines cohesion and separation. The similarity between objects and clusters is known as cohesion. A measure of how different objects are compared to other clusters is called splitting. This comparison is achieved by using Silhouette values between [-1, 1]. The high value in the Silhouette indicates that the objects have a close relationship with clusters. It is calculated by using the following steps [18]:

- First, find the average on other data in cluster *i* using the following equation:

$$a(i) = \frac{1}{n_i - 1}\sum_{q=1, q\neq i}^{n_i} d(x_i, x_q) \quad (5)$$

where:
$a(i)$ = average distance value
x = data
n_i = the number of data in the *i*-cluster, where *i*=1, 2, ... *k*

k = number of clusters
x_q = example of arbitrary data in cluster *i* except x_i

- Second, calculate the similarity between clusters by computing the average distance of the i-th data value from all points in the nearest cluster, then calculate the minimum as follows:

$$b(i) = \ min_{1 \le m \le k}(\frac{1}{n_m}\sum_{p=1}^{n_m} d(x_i, x_p)) \tag{6}$$

where:
$b(i)$ = minimum average value
x = data
n_m = the number of data in the *m*-cluster, where *i*=1, 2, ... *k*
k = number of clusters
x_p = example of arbitrary data in cluster *m* except x_i

- The coefficient of the silhouette of the point x_i is computed as follows:

$$si(i) = \frac{(b_i - a_i)}{\max\{(a_i, b_i)\}} \tag{7}$$

where:
b_i = the mean nearest cluster distance
a_i = the mean intra-cluster distance

- Finally, finding the mean *si*(*i*) of all data points is defined as the Silhouette Coefficient *i*.

$$S = \frac{1}{n}\sum_{i=1}^{n} si(i) \tag{8}$$

where:
S = the average value of $si(i)$
n = the number of data

Criteria of the Silhouette coefficient according to Kaufman and Rousseuw are described as follows:

Table 1. Silhouette coefficient value.

Silhouette Coefficient (SC)	Criteria
0.7 < SC <= 1	The strong structure
0.52 <= SC <= 0.7	The reasonable structure
0.26 <= SC <= 0.50	The weak structure
SC < 0.25	Not found a substantial structure

3. Result and Analysis

The trials in this study used 350 data. Each cluster has a silhouette score that is displayed in Figure 3. The X-axis describes the number of clusters. The Y axis expresses the silhouette score. Figure 3 describes that the best silhouette score of FCM was obtained from k = 2. The number of clusters and their silhouette score is reported in Table 2. The quality of clustering techniques is measured by using Silhouette Score. The following is information that described the meaning of Silhouette values that range from -1 to 1:

-1: Incorrect cluster used
0: The distance is not significant
1: Each cluster is separated

The high score in the silhouette represents the better grouping of objects into a cluster. A low score describes the worse data grouping in the cluster. The average value of each feature by cluster is reported in Table 3. The number of cluster distributions in cluster #0 is 183 while cluster #1 is 167. Table 3 and Table 4 show the average value for each parameter.

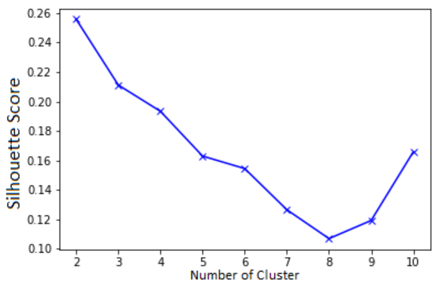

Figure 3(a). The silhouette score of each cluster with Euclidean distance

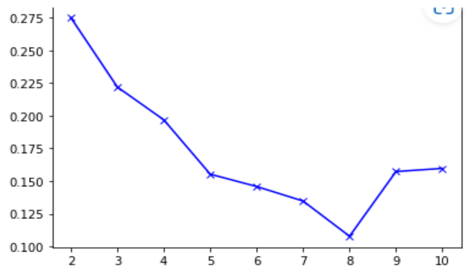

Figure 3(b). The silhouette score of each cluster with Manhattan distance

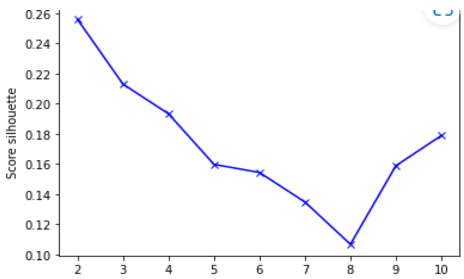

Figure 3(c). The silhouette score of each cluster with Minkowski

Table 2. Comparison of Silhouette score of Euclidian, Manhattan, and Minkowski

No.	No. of Cluster	Euclidean	Manhattan	Minkowski
0	2	0.255804	0.274880	0.255804
1	3	0.211137	0.221910	0.212911
2	4	0.193296	0.196834	0.193296
3	5	0.162889	0.155277	0.159789
4	6	0.154408	0.145849	0.154408
5	7	0.126435	0.134740	0.134858
6	8	0.106863	0.107830	0.106863
7	9	0.119164	0.157300	0.158875
8	10	0.155747	0.159676	0.178884

The average value of NaCl in cluster 0 is 94 while the average value of NaCl in cluster 1 is 93. Figure 4 displays a clustering plot of salt with k = 2 based on the distribution of each feature parameter. Some experiments have been carried out using the Silhouette Coefficient to find out the cluster structure. The results show that k = 2 with a Silhouette value of 0.255804, 0.274880, and 0.255804 by using Euclidean, Manhattan, and Minkowski respectively. From the test results, it is best to use Manhattan measurements. This shows that the clustering using the three distance measurement methods have the same structure by producing k=2.

Table 3. The average value of each feature by cluster.

CLUSTER	MOISTURE CONTENT	INSOLUBLE	CLAUSTROPHOBIA	MAGNESIUM	SULFATE	NaCl (Wb)	NaCl (db)
0	4.920967	0.444586	0.364493	0.351288	1.259038	89.880461	94.665187
1	0.151697	0.278084	0.317234	0.766164	1.220027	84.877398	93.383467

Graphs 3(a)-(c) show that there is a decrease in the Silhouette Coefficient value. This is because the values of each parameter are closer to k = 2. The following is a description of the parameter values for each cluster shown in Figure 4. It can be seen that the optimum value of the Fuzzy C-Mean cluster is highest at the cluster k = 2 while it reduces within range cluster 3 to 8. Optimum validity is in the lowest cluster which shows the distribution of data and its correlation value.

Table 4. The Characteristics of the Cluster

Characteristics	Cluster 0	Cluster 1
The score of moisture content	<= 0.5	<= 0.9
The score of insoluble	<= 0.4	<= 0.2
The score of claustrophobia	<= 3.6	<= 0.31
The score of magnesium	<= 0.3	<= 0.7
The score of sulfate	<= 1.25	<= 1.22
The score of NaCl(wb)	<= 89	<= 84
The score of NaCl(db)	<= 94	<= 93
Correlation of NaCl (wb) and NaCl (db)	NaCl(wb) has a positive correlation with NaCl(db), meaning that if the value of NaCl(wb) increases then the value of NaCl(db) also increases.	NaCl(wb) has a positive correlation with NaCl(db), meaning that if the value of NaCl(wb) increases, the value of NaCl(db) also increases.
Correlation of NaCl (wb), Moisture content and magnesium	NaCl(wb) has a negative correlation with Moisture content and magnesium, meaning that if the value of NaCl(wb) increases, the value of Moisture content and magnesium decreases.	NaCl(wb) has a negative correlation with Moisture content and magnesium, meaning that if the value of NaCl(wb) increases, the value of Moisture content and magnesium decreases.

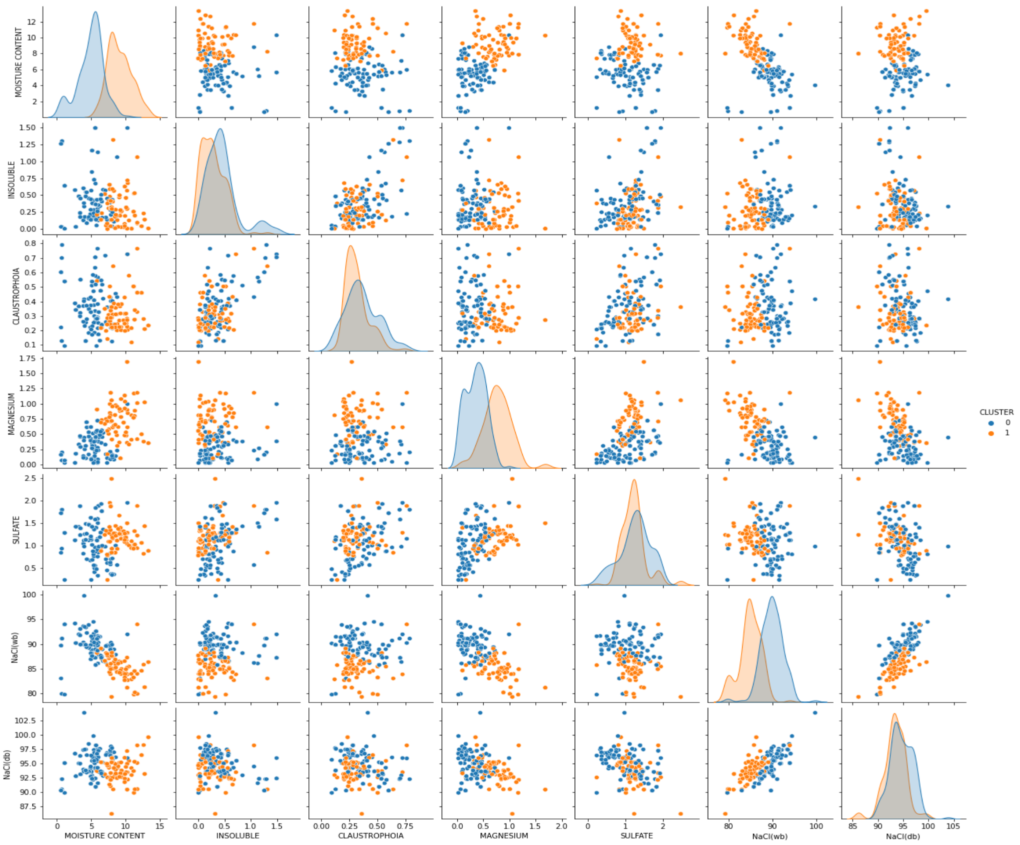

Figure 4. The plot of Clustering.

4. Conclusion

This research aims to exploit the Fuzzy C-Means Clustering as the method for grouping national salt commodities. This study used a test with the number of clusters k=10 for the distribution of training and testing data purposes. The mean imputation and the Fuzzy C-Means clustering method were utilized in the experimental research trials. The results show that the Fuzzy C-Means Clustering using the Silhouette Coefficient method with the calculation of the closest distance using Manhattan Distance is the best choice in this clustering research. The best result obtained in the Silhouette Coefficient assessment is at a value of 0.274880 with a value of k = 2.

5. Acknowledgment

We would like to thank LPPM - the University of Trunojoyo Madura for providing Research Funding through Independent Research-UTM in 2022 with contract number No. 225/UN46.4.1/ PT.01.03/2022

References

[1] Arif Haendra, M. Syamsul Maarif, Joko Affandi, Anggraini Sukmawati, Strategy to Increase the Competitiveness of National Salt in Indonesia, *Jurnal Manajemen & Agribisnis*, Vol. 18, No. 2, July 2021, doi: http://dx.doi.org/10.17358/jma.18.2.193.

[2] Chor-ching Goh, Combating iodine deficiency: Lessons from China, Indonesia, and Madagascar, *Food and Nutrition Bulletin*, Vol. 23, No. 3, 2002, The United Nations University, doi: 10.1177/156482650202300308.

[3] Richard D Semba, Saskia de Pee, Sonja Y Hess, Kai Sun, Mayang Sari, Martin W Bloem, Child malnutrition and mortality among families not utilizing adequately iodized salt in Indonesia, *The American Journal of Clinical Nutrition*, Volume 87, Issue 2, February 2008, Pages 438–444, https://doi.org/10.1093/ajcn/87.2.438.

[4] World Bank. Indonesia: Intensified Iodine Deficiency Control Project, *Implementation Completion Report*. Jakarta, Republic of Indonesia: World Bank, 2004.

[5] WHO/UNICEF/ICCIDD. *Assessment of Iodine Deficiency Disorders and Monitoring Their Elimination*. 3rd ed. France: World Health Organization; 2007. pp. 7-10.

[6] Rachmad Hidayat, Sabarudin Akhmad., Kukuh Winarso, Anis Arendra, K-Means Method for Determining Location of Facilities and Development of Supply Chain Network for Salt Commodities in Sumenep District, 2020 *Information Technology International Seminar* (ITIS). Surabaya, Indonesia, October 14-16, 2020.

[7] Ach. Khozaimi, Yoga Dwitya Pramudita, Eka Mala Sari Rochman, Aeri Rachmad, Salt Quality Determination Using Simple Additive Weighting (SAW) and Analytical Hirarki Process (AHP) Methods, *Kursor*, Vol. 10, Dec., 2019.

[8] Yeni Kustiyahningsih, Eza Rahmanita, Purbandini, Aeri Rachmad, Jaka Purnama, Integration interval type-2 FAHP-FTOPSIS group decision-making problems for salt farmer recommendation, *Commun. Math. Biol. Neurosci.* 2021.

[9] Ach. Khozaimi, Yoga Dwitya Pramudita, Eka Mala Sari Rochman, Aeri Rachmad, Decision Support System for Determining the Quality of Salt in Sumenep Madura-Indonesia, *ICComSET, Journal of Physics: Conference Series*, 1477, 2020.

[10] Deepak Sinwar, Rahul Kaushik." Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering". *International Journal For Research In Applied Science and Engineering Technology* (Ijras Et). Vol. 2 Issue V, May 2014.

[11] D T Pham, S S Dimov, and C D Nguyen, Selection of K in K-means clustering, *Proc. IMechE* Vol. 219 *Part C: J. Mechanical Engineering Science*.

[12] I.B.G.N. Giriputra, Missing value imputation using KNN method optimized with memetic algorithm, *e-Proc.Eng.* 3 (2016), 1098–1105.

[13] E. Acuña, C. Rodriguez, The treatment of missing values and its effect on classifier accuracy, in: D. Banks, F.R. McMorris, P. Arabie, W. Gaul (Eds.), *Classification, Clustering, and Data Mining Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004: pp. 639–647. https://doi.org/10.1007/978-3-642-17103-1_60.

[14] C.J. Tsai, C.I. Lee, W.P. Yang, A discretization algorithm based on class-attribute contingency coefficient, *Inform. Sci.* 178 (2008), 714–731. https://doi.org/10.1016/j.ins.2007.09.004.

[15] D. Borkin, A. Némethová, G. Michaľčonok, K. Maiorov, Impact of Data Normalization on Classification Model Accuracy, *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*. 27 (2019), 79–84. https://doi.org/10.2478/rput-2019-0029.

[16] R Suwanda, Z Syahputra and E M Zamzami, Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K, *ICCAI* 2019 *Journal of Physics: Conference Series*, 2020.

[17] Richa Loohach, Kanwal Garg, Effect of Distance Functions on K-Means Clustering Algorithm,. *International Journal of Computer Applications* (0975 – 8887) Volume 49– No.6, July 2012.

[18] Chunhui Yuan and Haitao Yang, Research on K-Value Selection Method of K-Means Clustering Algorithm, Graduate institute, *Space Engineering University*, Beijing 101400, China,. 2019.